Event-Enriched Image Analysis Grand Challenge at ACM Multimedia 2025

Thien-Phuc Tran*
University of Science, VNU-HCM, Vietnam

Minh-Quang Nguyen*
University of Science, VNU-HCM, Vietnam

Minh-Triet Tran
University of Science, VNU-HCM, Vietnam

Tam V. Nguyen
University of Dayton, Ohio, United States

Trong-Le Do
University of Science, VNU-HCM, Vietnam

Duy-Nam Ly
University of Science, VNU-HCM, Vietnam

Viet-Tham Huynh
University of Science, VNU-HCM, Vietnam

Khanh-Duy Le
University of Science, VNU-HCM, Vietnam

Mai-Khiem Tran
University of Science, VNU-HCM, Vietnam

Trung-Nghia Le†
University of Science, VNU-HCM, Vietnam

arXiv:2508.18904v1 [cs.CV] 26 Aug 2025

Abstract

The Event-Enriched Image Analysis (EVENTA) Grand Challenge, hosted at ACM Multimedia 2025, introduces the first large-scale benchmark for event-level multimodal understanding. Traditional captioning and retrieval tasks largely focus on surface-level recognition of people, objects, and scenes, often overlooking the contextual and semantic dimensions that define real-world events. EVENTA addresses this gap by integrating contextual, temporal, and semantic information to capture the who, when, where, what, and why behind an image. Built upon the OpenEvents V1 dataset, the challenge features two tracks: Event-Enriched Image Retrieval and Captioning, and Event-Based Image Retrieval. A total of 45 teams from six countries participated, with evaluation conducted through Public and Private Test phases to ensure fairness and reproducibility. The top three teams were invited to present their solutions at ACM Multimedia 2025. EVENTA establishes a foundation for context-aware, narrative-driven multimedia AI, with applications in journalism, media analysis, cultural archiving, and accessibility. Further details about the challenge are available at the official homepage: https://ltnghia.github.io/eventa/eventa-2025.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Computer vision**; **Machine learning**.

Keywords

Event-enriched image analysis, article retrieval, image retrieval, image captioning

*Both authors contributed equally to this research.
†Corresponding author. Email: ltnghia@fit.hcmus.edu.vn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762067

ACM Reference Format:
Thien-Phuc Tran, Minh-Quang Nguyen, Minh-Triet Tran, Tam V. Nguyen, Trong-Le Do, Duy-Nam Ly, Viet-Tham Huynh, Khanh-Duy Le, Mai-Khiem Tran, and Trung-Nghia Le. 2025. Event-Enriched Image Analysis Grand Challenge at ACM Multimedia 2025. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746027.3762067

1 Introduction

Recent advances in multimodal AI have significantly improved image captioning [17], object detection [3], and semantic segmentation [10], enabling systems to describe visible content with high accuracy. Nevertheless, these methods largely remain limited to surface-level recognition of people, objects, and actions [2]. In real-world applications, such literal descriptions are insufficient. For example, identifying that "people are watching television" leaves critical contextual questions unanswered: what program are they watching, is it culturally or politically significant, and why are they enjoying it? Current AI systems rarely address such deeper dimensions of meaning, resulting in captions that lack explanatory depth and narrative coherence [6].

The Event-Enriched Image Analysis (EVENTA) Grand Challenge addresses this gap by shifting the focus from visual analysis to event-level understanding. Its central objective is to transform image interpretation into contextualized storytelling that explains not only what is seen but also who is involved, when and where the event takes place, what is happening, and why it is significant. Achieving this requires combining visual information with contextual reasoning, thereby producing narrative-rich captions that situate images within broader social, historical, or cultural frames, thus providing the audience with deeper insights into the semantics underlying the perceived images and textual content.

By introducing this event-aware perspective, our EVENTA challenge advances a model of perception that captures not only visual elements but also their contextual significance. Understanding an image becomes an exercise in reconstructing its context, implications, and human relevance, rather than merely enumerating objects and scenes. For instance, a protest in a city square, a historic moment captured in a news photograph, or a family gathering suffused with subtle emotions each demands interpretations that extend beyond literal recognition. Accordingly, our challenge highlights the

need for richer semantic descriptions that capture not only the visible signals but also their deeper event-centric meaning. The implications of this paradigm extend across multiple domains. In journalism and media analysis, event-enriched captioning can support automatic generation of accurate and informative descriptions for archival and contemporary news imagery. For event discovery and retrieval, the integration of contextual knowledge enables more expressive and semantically meaningful search capabilities. In cultural archiving and storytelling, such methods provide tools for preserving socio-historical moments with narrative fidelity.

To this end, the EVENTA 2025 Grand Challenge promotes the development of context-aware captions that incorporate event-specific details such as names, timelines, and outcomes. By providing a benchmark that evaluates semantic depth, our challenge offers the community a testbed for advancing multimodal systems beyond descriptive accuracy toward genuine contextual understanding. The EVENTA challenge attracted broad international participation, with 45 teams from six countries registering across its two tracks: Event-Enriched Image Retrieval and Captioning and Event-Based Image Retrieval. The evaluation was conducted in two stages: a Public Test phase for system development and tuning, followed by a Private Test phase to ensure fair and unbiased assessment. Based on the final leaderboard, the top three teams were invited to present their solutions at ACM Multimedia 2025, recognizing their outstanding contributions to advancing event-aware multimodal understanding.

The success of EVENTA 2025 demonstrates not only the feasibility of integrating vision, retrieval, and contextual reasoning but also the growing interest of the community in moving beyond surface-level recognition. By presenting a benchmark for event-enriched captioning and retrieval, EVENTA offers a basis for subsequent studies that connect computer vision, natural language processing, and information retrieval. The results indicate the potential of context-rich, interpretive approaches to visual media and open avenues for developing AI systems oriented toward understanding not just what is shown, but also who is involved, when and where it occurs, and why it matters.

2 Related Work

2.1 Image Captioning

Recent advances in image captioning leverage large-scale vision–language pretraining and unified architectures. Contrastive models like CLIP demonstrated the power of aligning visual and textual embeddings on web-scale data, enabling strong zero-shot recognition and retrieval [13]. Building on this, transformer-based frameworks (e.g., ALBEF) first align image–text representations and then fuse them with cross-modal attention; its align-before-fuse strategy with momentum distillation achieved state-of-the-art results on captioning and multimodal understanding [8]. BLIP unified vision–language understanding and generation by bootstrapping captions for noisy web images through a captioner-and-filter scheme and a flexible encoder–decoder, improving COCO retrieval and captioning benchmarks [7]. Generative transformer models have also emerged: GIT simplifies the architecture to one ViT encoder and one text decoder under a language-modeling objective, scaling pretraining to surpass human-level CIDEr on TextCaps [16].

Another line of work focuses on multimodal reasoning and external knowledge to produce richer, context-aware captions. Flamingo bridges pretrained visual encoders and large language models via novel cross-attention mechanisms, enabling in-context learning for open-ended vision–language tasks without task-specific fine-tuning [1]. Retrieval-augmented captioning injects real-world context: EXTRA conditions a captioner on top-k captions retrieved from a datastore and encodes image + retrieved text with a V&L encoder, significantly improving quality on COCO [14]. In the news domain, Visually-Aware Context Modeling integrates article signals (e.g., face–name cues) to guide caption generation, enabling naming and event details beyond the visible content [12]. EVCap augments LLM-based captioning with an external visual–name memory to improve open-world comprehension and novel-object grounding [9]. Despite progress, models often miss event-level semantics (specific names, temporal context, implications) not directly visible, motivating event-grounded captioning that goes beyond surface descriptions.

2.2 Image Retrieval

Image–text retrieval has similarly advanced via powerful vision–language representations and large datasets. Dual-encoder models trained on millions to billions of pairs (e.g., CLIP, ALIGN) learn a shared embedding space that enables effective cross-modal retrieval, even zero-shot [5, 13]. Subsequent methods improved fine-grained alignment: ALBEF's contrastive align-before-fuse pretraining yields strong retrieval with modest data [8], and BLIP's caption filtering further boosts COCO recall [7]. More recently, ALIP integrates synthetic captions to mitigate web noise, achieving state-of-the-art zero-shot image–text retrieval [18].

Beyond generic short-query retrieval, event-centric queries require aligning narrative descriptions with visual scenes. Systems increasingly consider contextual cues (articles, names, timelines) and leverage language models to parse complex queries; yet long, narrative queries remain challenging, highlighting the need for event-grounded retrieval capable of bridging vision and storytelling—precisely the focus of EVENTA-style tasks [9, 12, 14].

Motivated by the gaps in current multimodal understanding research, we organize the EVENTA Grand Challenge at ACM Multimedia 2025. The tasks are designed to encourage participants to explore deeper layers of meaning beyond visible content, underscoring the need for event-centric understanding of multimedia and fostering the development of systems for event-based downstream tasks.

3 Challenge Description

3.1 Tasks

3.1.1 Track 1: Event-Enriched Image Retrieval and Captioning. This track is designed to generate captions that deliver richer and more comprehensive accounts of an image (Figure 1). Unlike conventional captioning tasks that focus solely on surface-level visual elements, the objective here is to produce context-aware descriptions that incorporate entity names, attributes, temporal and spatial information, event outcomes, and other details that cannot be inferred from visual content alone. Given an input image, participants are required to retrieve relevant evidence from a curated external article

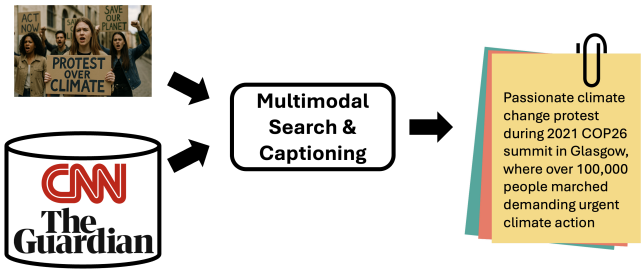


Figure 1: Given a dataset of news, including images and content, with a set of query images. The task is to utilize multimodal search for retrieving related figures and apply captioning models to generate descriptions for each query.

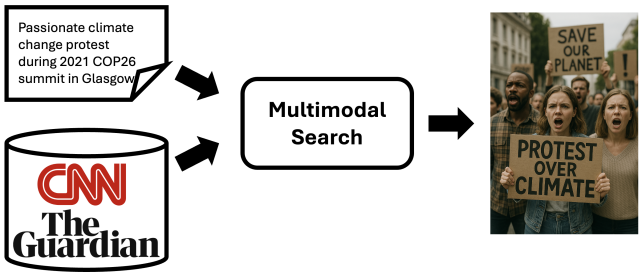


Figure 2: This task involves mainly image retrieval by browsing the most related image from the dataset for each query description, given a dataset of news, including images and content.

database and integrate this information into the caption generation process. By adopting a retrieval-augmented generation (RAG) paradigm, this track promotes the creation of coherent, narrative-driven captions that capture not only the visible elements of a scene but also its broader context and significance. The resulting captions therefore move beyond description toward event-level understanding, offering a more complete and semantically grounded interpretation of what the image represents.

3.1.2 Track 2: Event-Based Image Retrieval. In this track, participants are provided with a realistic event-related caption and are required to retrieve the corresponding images from a curated database (Figure 2). This task represents a fundamental problem at the intersection of computer vision and natural language processing, requiring the construction of a joint representation space in which visual and textual modalities can be effectively aligned and compared. Text-based image retrieval has broad applications, including search engines, medical imaging, e-commerce, and digital asset management. Despite substantial progress, several challenges remain, such as addressing abstract or ambiguous queries, scaling retrieval systems to large datasets, and ensuring robustness to linguistic variability and biases in training data. By situating retrieval within the context of real-world event information, this track emphasizes the need for models that can move beyond surface-level matching toward event-aware retrieval, thereby advancing both the accuracy and contextual relevance of cross-modal search.

Table 1: Evaluation metrics used in Track 1: Event-Enriched Image Retrieval and Image Captioning.

Metrics	Description	Domain
AP	Average Precision—measures retrieval precision across thresholds	Retrieval
Recall@1	Whether the correct item is ranked at the top	Retrieval
Recall@10	Whether the correct item is within the top 10 results	Retrieval
CLIPScore [4]	Semantic alignment between image and caption	Captioning
CIDEr [15]	Agreement between generated and reference captions	Captioning

Table 2: Evaluation metrics used in Track 2: Event-Based Image Retrieval.

Metric	Description
mAP	Mean Average Precision—measures overall retrieval precision
MRR	Mean Reciprocal Rank—measures how early the first correct item is retrieved
Recall@1	Whether the correct item is ranked first
Recall@5	Whether the correct item is within the top 5 results
Recall@10	Whether the correct item is within the top 10 results

3.2 Evaluation Metrics

For both tasks, we propose and select carefully a variation of metrics for the overall evaluation and fair performance ranking between each team. To fairly and comprehensively evaluate submissions in both tracks of our challenge, we propose a combined score calculation protocol by assigning a weight tensor $\mathbf{w}$ to the component score and computing the **weighted harmonic mean** of the score as the final overall score for each valid submission. The formula to compute the overall score is provided as follows:

$$\texttt{Overall Score} = r \cdot \frac{\sum_i w_i}{\left(\sum_i \frac{w_i}{\texttt{score}_i + \epsilon}\right)}, \tag{1}$$

where $\epsilon = 10^{-5}$ is a predefined constant to avoid diminishing denominator during the calculation, and r is the number of valid responses in the submission.

3.2.1 Track 1: Event-Enriched Image Retrieval and Captioning. For track 1, we employed two separate sets of metrics classified based on their specific domains - Image Retrieval and Image Captioning. The metrics are provided as in Table 1. The overall score of submissions are calculated based on Equation 1. After careful experiments and overall evaluation under different settings, we chose the set of weights

$$\begin{aligned}\mathbf{w} &= [w_{\texttt{AP}},\ w_{\texttt{R@1}},\ w_{\texttt{R@10}},\ w_{\texttt{CLIPScore}},\ w_{\texttt{CIDEr}}] \\ &= [0.1,\ 0.2,\ 0.2,\ 0.3,\ 0.2]\end{aligned}$$

3.2.2 Track 2: Event-Based Image Retrieval. In track 2, we chose five metrics to compute the retrieval scores in a similar fashion to track 1. The chosen metrics are described in Table 2. After careful

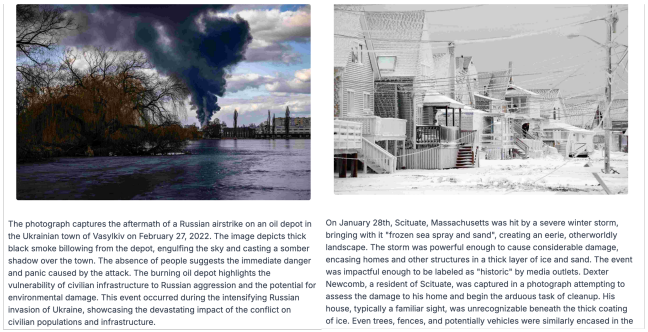


Figure 3: Samples in OpenEvents V1 dataset [11].



Figure 4: Image query in the Private Set of Track 1.

analyses, we pre-determined the set of weights for this track to be

$$\mathbf{w} = [w_{\text{mAP}}, w_{\text{MRR}}, w_{\text{R@1}}, w_{\text{R@5}}, w_{\text{R@10}}] = [0.3,\ 0.2,\ 0.2,\ 0.15,\ 0.15]$$

3.3 Dataset

The EVENTA 2025 Grand Challenge utilizes the OpenEvents V1 dataset [11], a large-scale, event-driven corpus designed to bridge the gap between visual content and real-world news understanding. Collected from more than a decade of reporting by two major international outlets, CNN and The Guardian, the dataset captures the dynamic intersection of images, events, and storytelling (Figure 3). It comprises over 200,000 news articles paired with more than 400,000 images, spanning the years 2011–2022 and covering diverse domains such as politics, climate, technology, culture, and sports. To support rigorous evaluation, the dataset also includes more than 30,000 annotated image–event caption pairs, curated and organized into training, public test, and private test splits. This combination of scale, diversity, and event-centric annotation provides a robust foundation for benchmarking models that seek to advance event-enriched image captioning and retrieval. We remark that image queries in the Private Set of Track 1 were heavily augmented to simulate various scenarios in the real-life (Figure 4).

3.4 Rules

The EVENTA challenge is conducted under the following rules to ensure fairness, reproducibility, and scientific integrity:

- Participants are not allowed to annotate the test sets.
- External datasets are permitted. However, participants may only use publicly accessible datasets and pre-trained models. The use of private datasets or pre-trained models is strictly prohibited.
- Open sources are allowed. However, commercial tools, libraries, APIs, etc. are strictly prohibited.
- Final scores are calculated based on performance on the Private Test set.
- Participants must make their source code publicly available on GitHub to ensure reproducibility.
- Participants must submit a detailed paper through the official challenge platform before the deadline to validate their solutions.
- Late submissions would not be accepted.
- Only registered teams that submit papers are eligible to win, but all participants' scores are recognized.

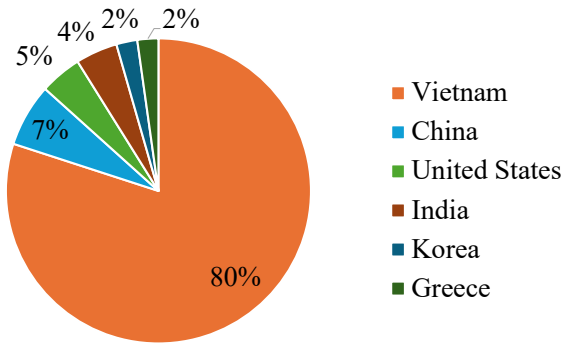


Figure 5: Demographics of participating teams in the EVENTA 2025 Grand Challenge.

3.5 Participants

The challenge attracted 45 registered teams representing six countries. The majority of participants came from Vietnam, with 36 teams from institutions including the University of Science, the University of Information Technology, Ha Noi University of Science and Technology, and Science Lab. Additional participation included one team from Greece (National Technical University of Athens), two teams from the United States (Northwestern University and Hitachi America Research & Development), three teams from China (Harbin Institute of Technology, Fudan University, and South China University of Technology), one team from Korea (Kookmin University), and two teams from India (Jaypee Institute of Information Technology and Indian Institute of Technology Dhanbad). This diverse participation highlights the strong international interest in advancing event-enriched multimedia understanding.

4 Challenge Results

4.1 Track 1: Event-Enriched Image Retrieval and Captioning

Track 1 evaluated approaches on retrieving relevant contextual evidence and generating event-enriched captions. Most teams converged on a retrieval-augmented captioning paradigm, where external article evidence was first retrieved and re-ranked, then fused with visual features for caption generation. Common design traits

Table 3: Performance of top teams in the Private Set of Track 1.

Rank	Team	Overall	AP	R@1	R@10	CLIPScore	CIDEr
1	Cerebro	0.550	0.991	0.989	0.995	0.826	0.210
2	SodaBread	0.547	0.982	0.977	0.988	0.870	0.205
3	Re:zero Slavery	0.451	0.955	0.945	0.973	0.732	0.156
4	ITxTK9	0.420	0.966	0.955	0.983	0.828	0.133
5	HCMUS-NoName	0.282	0.708	0.663	0.801	0.783	0.081

Table 4: Performance of top teams in the Private Set of Track 2.

Rank	Team	Overall	mAP	MRR	R@1	R@5	R@10
1	NoResources	0.577	0.563	0.563	0.469	0.690	0.744
2	23Trinitrotoluen	0.572	0.558	0.558	0.456	0.698	0.762
3	LastSong	0.563	0.549	0.549	0.449	0.695	0.738
4	Sharingan Retrievers	0.533	0.521	0.521	0.428	0.640	0.705
5	ZJH-FDU	0.368	0.361	0.361	0.270	0.491	0.525

included the use of strong vision–language embeddings for candidate retrieval, multi-source contextual signals to enrich captions, and large language models guided by structured prompting. While these approaches achieved near-perfect retrieval accuracy (AP and Recall@1 above 0.95 for the top teams), CIDEr scores remained modest, revealing that fluent and contextually faithful caption generation is still the bottleneck (Table 3).

Among the top performers, **Cerebro** secured first place with an overall score of 0.550. Its technical strength lies in uncertainty-guided re-ranking, using entropy, distance, and similarity heuristics, to refine candidate selection, coupled with chain-of-thought prompting that integrates five contextual sources (article summaries, QA pairs, named entities, generated captions, and crawled captions). This comprehensive reasoning pipeline enabled narrative-rich captions and near-perfect retrieval (AP = 0.991, Recall@1 = 0.989). **SodaBread**, a close second with 0.547, emphasized precision at both retrieval and captioning stages by pairing DINOv2 global similarity with patch-level nearest neighbor matching, followed by multi-source context fusion in Qwen3. Its Semantic Gaussian Normalization module further improved factual grounding and fluency, yielding the highest CLIPScore (0.870) among all teams. **Re:zero Slavery**, ranked third with 0.451, combined BEiT-3 and SigLIP retrieval with ORB and SIFT re-ranking, and used InstructBLIP plus a fine-tuned Qwen3 (via QLoRA) to generate paragraph-level enriched captions that strongly emphasized temporal cues and named entities.

4.2 Track 2: Event-Based Image Retrieval

Track 2 focused on retrieving the most relevant images given realistic event-related captions, a task that requires bridging complex textual event descriptions with visually grounded evidence. All leading teams adopted multi-stage approaches that share several common design principles: (1) an initial article retrieval step using strong pretrained language or multimodal embeddings to narrow the candidate pool, (2) a reranking stage to refine results using additional semantic or contextual signals, and (3) an image-level matching step that integrates vision–language models for cross-modal alignment. Moreover, most approaches employed fusion strategies such as Reciprocal Rank Fusion (RRF) to combine outputs from multiple retrieval signals, highlighting a trend toward robustness through ensemble-style aggregation. While Recall@10 exceeded 0.70 for the top teams, Recall@1 remained relatively low, underscoring the persistent challenge of fine-grained semantic grounding between narrative-rich captions and precise visual matches (Table 4).

Despite these shared strategies, the top three teams distinguished themselves through their specific techniques. **NoResources**, the winning system, leveraged a multi-stage framework where Qwen3-embedding handled article retrieval, Qwen3-reranker performed event-aware reranking, and Qwen2-VL scored caption–image similarity, with RRF providing robustness. This design yielded the highest overall score (0.577), demonstrating strong adaptability and balanced performance across all metrics. **23Trinitrotoluen**, ranked second with a score of 0.572, introduced a hybrid dense–sparse strategy by integrating OpenCLIP, Nomic embeddings, and BM25 retrieval, enhanced with semantic chunking and boosting for reranking. This balance of semantic richness and lexical precision proved highly effective in large-scale event retrieval. **LastSong**, finishing third with 0.563, emphasized hierarchical text modeling by combining sparse, chunked, long-context, and document-level embeddings for article retrieval, followed by reranking and vision-language alignment for image matching. Its fusion of article and image similarity enabled competitive performance, with only a narrow margin separating it from the top two teams.

Overall, the results across both tracks demonstrate the feasibility of event-enriched captioning and retrieval while also revealing ongoing challenges. Top-performing teams leveraged retrieval-augmented strategies and multimodal fusion to achieve strong performance, but gaps remain in generating fluent, contextually accurate captions and in handling fine-grained event-level retrieval. These findings highlight promising directions for future research, including improved integration of external knowledge sources, robust evaluation metrics, and models capable of reasoning over abstract or ambiguous event queries.

5 Research Impact

The EVENTA 2025 Grand Challenge establishes a novel research agenda in multimedia understanding by advancing from surface-level description toward event-enriched interpretation of visual content. Unlike traditional captioning and retrieval tasks that emphasize objects, scenes, or actions in isolation, EVENTA explicitly integrates contextual dimensions such as participants, temporal and spatial information, outcomes, and significance. This event-driven perspective positions the challenge at the intersection of computer vision, natural language processing, and information retrieval, encouraging the development of models that can generate narrative-rich and semantically grounded outputs.

EVENTA Challenge also provides the research community with a platform to explore new approaches for RAG, multimodal fusion, and knowledge-grounded captioning. By prioritizing factual accuracy, contextual depth, and narrative coherence, the challenge also promotes evaluation protocols that move beyond conventional automatic metrics, driving the community toward more comprehensive and human-centered assessments of multimodal understanding.

The broader significance of our EVENTA Challenge extends to multiple real-world applications, including journalism, digital

archiving, media search, accessibility, and cultural preservation. By demonstrating the feasibility of systems capable of answering the fundamental who, when, where, what, and why questions of an image, the challenge lays the foundation for next-generation multimedia technologies that are more context-aware, interpretable, and socially relevant. Through strong international participation and interdisciplinary engagement, we foster a vibrant community of researchers and practitioners dedicated to advancing event-driven multimedia AI.

6 Conclusion and Future Directions

The EVENTA 2025 Grand Challenge introduced the first large-scale benchmark for event-enriched image captioning and retrieval, addressing the limitations of conventional visual understanding by explicitly emphasizing contextual, narrative, and event-level semantics. With participation from 45 teams across six countries, the challenge demonstrated both the feasibility of new tasks and the strong global interest in advancing event-aware multimodal intelligence. Participant submissions underscored the potential of RAG and multimodal fusion, while also highlighting important challenges that remain related to factual grounding, narrative coherence, and generalization across diverse event domains.

EVENTA opens several promising avenues for future research. Expanding the dataset to incorporate additional modalities such as video, audio, or social media streams would enable richer and more dynamic event understanding. Developing evaluation protocols that better capture factual accuracy, contextual depth, and narrative quality remains an important direction, as current metrics often fail to reflect human judgments of informativeness and coherence. Furthermore, the integration of structured knowledge bases, temporal reasoning, and causal inference has the potential to enhance system performance in capturing the who, when, where, what, and why dimensions of events. Finally, continued collaboration across computer vision, natural language processing, information retrieval, and the digital humanities will be essential to advancing the broader research agenda of context-aware, event-driven multimedia AI.

Through EVENTA 2025 Grand Challenge, we have taken a decisive step toward bridging the gap between vision and storytelling. We hope this challenge will inspire the community to pursue the development of systems that move beyond describing what is visible to interpreting the meaning, implications, and human significance of visual media.

Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

We would like to thank the ACM Multimedia 2025 organizers for agreeing to host our EVENTA Challenge and for their support. We also thank the Codabench open source competition platform for hosting our leaderboard submissions.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

[2] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12466–12475.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139.

[6] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 317–325.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 12888–12900.

[8] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*.

[9] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024. EVCap: Retrieval-Augmented Image Captioning with External Visual–Name Memory for Open-World Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[11] Hieu Nguyen, Phuc-Tan Nguyen, Thien-Phuc Tran, Minh-Quang Nguyen, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2025. OpenEvents V1: Large-Scale Benchmark Dataset for Multimodal Event Grounding. In *ACM International Conference on Multimedia*.

[12] Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2024. Visually-Aware Context Modeling for News Image Captioning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 8748–8763.

[14] Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-Augmented Image Captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 3648–3663.

[15] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv:2205.14100* (2022).

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[18] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. ALIP: Adaptive Language-Image Pre-Training with Synthetic Caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.